



GLAWI, a free XML-encoded Machine-Readable Dictionary built from the French Wiktionary

Franck Sajous, Nabil Hathout

► To cite this version:

Franck Sajous, Nabil Hathout. GLAWI, a free XML-encoded Machine-Readable Dictionary built from the French Wiktionary. eLex, Aug 2015, Herstmonceux, United Kingdom. halshs-01191012

HAL Id: halshs-01191012

<https://shs.hal.science/halshs-01191012>

Submitted on 1 Sep 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

GLAWI, a free XML-encoded Machine-Readable Dictionary built from the French Wiktionary

Franck Sajous and Nabil Hathout

CLLE-ERSS (CNRS & Université de Toulouse 2)
franck.sajous@univ-tlse2.fr, nabil.hathout@univ-tlse2.fr

Abstract

This article introduces GLAWI, a large XML-encoded machine-readable dictionary automatically extracted from Wiktionnaire, the French edition of Wiktionary. GLAWI contains 1,341,410 articles and is released under a free license. Besides the size of its headword list, GLAWI inherits from Wiktionnaire its original macrostructure and the richness of its lexicographic descriptions: articles contain etymologies, definitions, usage examples, inflectional paradigms, lexical relations and phonemic transcriptions. The paper first gives some insights on the nature and content of Wiktionnaire, with a particular focus on its encoding format, before presenting our approach, the standardization of its microstructure and the conversion into XML. First intended to meet NLP needs, GLAWI has been used to create a number of customized lexicons dedicated to specific uses including linguistic description and psycholinguistics. The main one is GLÀFF, a large inflectional and phonological lexicon of French. We show that many more specific on demand lexicons can be easily derived from the large body of lexical knowledge encoded in GLAWI.

Keywords: French Machine-Readable Dictionary; Free Lexical Resource; Wiktionary; Wiktionnaire

1. Introduction

Recent papers on electronic lexicography investigate if and how linguistics (computational or not) can contribute to lexicography (Rundell, 2012), how NLP can automate the process of collecting material and analyze it (Rundell and Kilgarriff, 2011) or what are the skills and the needs of specific end-users (Lew, 2013). As linguists and NLP researchers, we are reciprocally interested in the exploitation of dictionaries for linguistic description (phonology, morphology, lexicology, semantics, etc.) and NLP use. Leveraging machine-readable dictionaries (MRDs) for the acquisition of lexical and semantic relations, for the development of derived lexical resources, or for various linguistic studies, was common practice in 1980's (Chodorow et al., 1985; Markowitz et al., 1986; Calzolari, 1988). The availability of large corpora and the subsequent rise of corpus linguistics highlighted MRDs' restricted coverage and their potential out-of-dateness. However, new online dictionaries with no size restriction and a steadily ongoing development such as Wiktionary may renew the interest for electronic lexicons. Besides its wide coverage and its

potential for constant updates, Wiktionary has an interesting macrostructure and features a rich lexical knowledge: articles include etymologies, definitions, lemmas and inflected forms, lexical semantic and morphological relations, translations and phonemic transcriptions.

For six years, we have exploited Wiktionary and more specifically its French language edition called Wiktionnaire, assessed its quality and investigated to what extent it can meet linguistics and NLP's needs in terms of lexical resources. Each experiment led us to extract various information from the collaborative dictionary and develop specific resources targeting different uses. In order to experiment algorithms based on random walks to enrich lexical networks (Sajous et al., 2010), we produced partial XML versions of the French and the English editions of Wiktionary, called WiktionaryX.¹ This resource contains a selection of fields extracted from the English and French wiktionaries: definitions, lexical semantic relations and translations. We then produced an inflectional lexicon called GLÀFF (Sajous et al., 2013a; Hathout et al., 2014b) that contains inflected forms, lemmas, morphosyntactic features and phonemic transcriptions.² This lexicon was intended to be used by syntactic parsers like Talismane (Urieli, 2013) or for research in computational morphology (Hathout, 2011; Hathout and Namer, 2014). A conclusion we drew is that Wiktionnaire's rich content is a valuable resource whose main drawback is its heterogeneous and volatile format, which impedes an easy and direct exploitation. A significant contribution of GLAWI is the standardization of Wiktionnaire's microstructure. Standing for "*GLÀFF and WiktionaryX*", GLAWI also results from our will to unify parallel efforts and produce a single resource that includes all information contained in Wiktionnaire in a workable format (XML). It is however not a simple merge of GLÀFF and WiktionaryX: new information is also extracted, like the morphological relations omitted from the two previous resources. We also went one step further in the homogenizing process. Our aim is to finely parse Wiktionnaire so that we can make accessible in a standard and coherent format as much information as available. To that extent, our approach differs from that of Sérasset (2012), whose aim is to build a multilingual network containing "easily extractable" (i.e. *regular*) entries, which results in a restricted coverage. Conversely, we made a particular effort to detect information, whatever format it is encoded into and wherever it occurs.

GLAWI is conceived as a general-purpose MRD intended to be easy to use, like such or as a starting point to tailor specific lexicons. GLÀFF, as well as other resources that we extracted so far from Wiktionnaire, will now be derived easily from GLAWI.

This article is organized as follow: in section 2, we give some insights into the Wiktionnaire's nature ; we describe GLAWI in section 3 and explain how we developed it by converting Wiktionnaire into a structured format. We illustrate in section 4 how we derived specific lexicons for various purposes directly from GLAWI, before contemplating some perspectives in section 5.

¹WiktionaryX is available at http://redac.univ-tlse2.fr/lexicons/wiktionaryx_en.html

²GLÀFF is available at http://redac.univ-tlse2.fr/lexicons/glaff_en.html

2. Wiktionary and Wiktionnaire

Wiktionary, presented as “*the lexical companion to Wikipedia*”,³ is, as Wikipedia and other related wikis, a public collaborative project. Any internet user can contribute, whatever their skills. Editorial policies exist, however modifications are published immediately. “Wiktionary” is used to refer both to the English edition and to the whole project (the 171 language editions). We hereafter give some details about the nature of Wiktionary and its French edition called Wiktionnaire.⁴

General description. The basic unit of Wiktionnaire’s articles is the word form. A given article (described in a Web page, at a URL) may contain several entries having distinct or identical parts of speech (POSs). A POS section may correspond to a canonical form (lemma) or an inflected form. Figure 1a depicts an excerpt of the page of *affluent*.

This page shows that the word form is the lemma of an adjective ‘tributary’, a noun ‘tributary’, and is an inflected form of the verb *affluer* ‘to flow’. The adjective POS-section gives the four inflected forms of its paradigm, each form linking to a dedicated page of the dictionary. Figure 1c shows the page corresponding to the feminine singular form *affluente*, which links back to the lemmatized form *affluent*. The inflected verbal forms of Figure 1a link to the page of the infinitive form, depicted in Figure 2. Unlike the pages of noun and adjective lemmas, the ones corresponding to verb infinitive forms do not contain their paradigms (a verb’s paradigm amounts to 48 forms in French which would cause a display overload). Instead, a link to a conjugation table is inserted. A shortened example of such a table is given for *affluer* in Figure 3. Each inflected form links to a dedicated page, when this page exists. This hypertextual macrostructure shows that the relations between the different forms of a given paradigm are located in different parts of the dictionary. We discuss the incidence of this feature in section 3.2.

The microstructure of an article contains an etymology section and one or more POS sections which provide a sense inventory including glosses and examples. POS sections may also include translations, lexical semantic relations (synonymy / antonymy, hypernymy / hyponymy, holonymy / meronymy), morphological relations (derivation, compounds) or more fuzzy relations such as *apparentés* ‘related’. Phonemic transcriptions may appear at the article level (when all entries share a common pronunciation), in the first line of the POS level and/or in the paradigms. It is worth noting that each language edition has its own microstructure. For example, the semantic relations are indexed to the word senses in the German Wiktionary. They are listed in POS sections in Wiktionnaire but appear at the article top level in the Italian Wiktionary.


An inappropriate software infrastructure (and its consequences). Launched in 2003, one year after the English edition, Wiktionnaire’s underlying infrastructure is the MediaWiki engine, used

³<http://en.wiktionary.org>

⁴Additional descriptions can be found in (Navarro et al., 2009; Meyer, 2013; Sajous et al., 2013b)

<http://fr.wiktionary.org/wiki/affluent>

affluent




Adjectif

affluent

1. (*Géographie*) Qui se **jette** dans un **autre** en **parlant** d'un **cours** d'eau.

2. (*Médecine*) Qui **affluent**, qui se **portent** en **abondance** vers **quelque** **partie** du **corps**.




Nom commun

affluent

/a.fly.ɑ̃/

masculin

1. (*Géographie*) **Cours** d'eau qui se **jette** dans un **autre**.



Forme de verbe

affluent

/a.fly/

1. *Troisième* *personne* du *pluriel* de l'*indicatif* *présent* de **affluer**.

2. *Troisième* *personne* du *pluriel* du *subjonctif* *présent* de **affluer**.

Singulier

Pluriel

Masculin

affluent

/a.fly.ɑ̃/

Féminin

affluente

/a.fly.ɑ̃t/

Singulier

Pluriel

affluent

affluents

/a.fly.ɑ̃/

INDICATIF

Présent

ils/elles affluent

SUBJONCTIF

Présent

qu'ils/elles affluent

(a) POS sections of the article *affluent*

```



{{-adj-|fr}}
{{fr-accord-cons|a.fly.ɑ̃|t}}
'''affluent'''
# {{géographie|fr}} Qui se [[jeter|jette]] [[dans]] un [[autre]] en [[parlant]]
  d'un [[cours]] d'eau.

{{-nom-|fr}}
{{fr-rég|a.fly.ɑ̃}}

{{-flex-verb-|fr}}
{{fr-verbe-flexion|affluer|ind.p.3p=oui|sub.p.3p=oui|}}
'''affluent''' {{pron|a.fly|fr}}
# '''Troisième personne du pluriel de l'indicatif présent de''' [[affluer]].
# '''Troisième personne du pluriel du subjonctif présent de''' [[affluer]].

```

(b) Wikicode of the article *affluent*

http://fr.wiktionary.org/wiki/affluente	<code>{{-flex-adj- fr}}</code>
 affluente	<code>'''affluente''' {{f}} {{pron a.fly.ɑ̃t}}</code>
 Forme d'adjectif	<code># '''Féminin singulier de''' [[affluent]].</code>
affluente <i>féminin</i> <i>/a.fly.ɑ̃t/</i>	
1. <i>Féminin singulier de affluent.</i>	

(c) Article *affluente* and corresponding wikicode

Fig. 1: Excerpts of Wiktionnaire's articles *affluent* and *affluente* and their underlying wikicode.

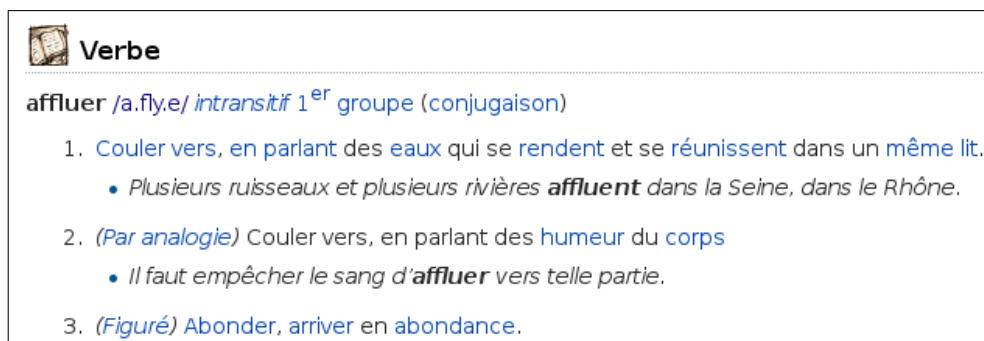


Fig. 2: Excerpt of Wiktionnaire's article *affluer*

by all the Wikimedia projects. Examples of the encoding format, called *wikicode*, are given in Figures 1b and 1c.

Rundell and Kilgarrieff (2011) attribute to Laurence Urdang the first vision, in mid 1960's, of the dictionary as a database “*facilitating and rationalizing the capture, storage and manipulation of dictionary text*”. Systematic check of cross-references was seen as an early benefit of this approach. Four decades later, Wiktionary, a dictionary born online, was encoded into unstructured text, ignoring the necessity of a database oriented design. Evan Jones, the author of the tool *wikipedia2text*,⁵ states that “*one of the biggest problems is that there is no well-defined parser for the wiki text that is used to write the articles. The parser is a mess of regular expressions, and users frequently add fragments of arbitrary HTML*”. Several consequences arise from this situation:

1. as no formal syntax of the wikicode is defined, no compliance-check is performed when a contributor edits an article. Encoding errors add to occasional contributors' amateurism.
2. cross-references and consistency checking is impossible. For example, a possible discrepancy between an inflected form given in its dedicated page and another form given in its lemma's paradigm cannot be detected. Similarly, Figure 1b shows that the same information, namely the inflectional features of the verbal form, appears in two ways: *affluent* as third person plural indicative of *affluer* is both given by the code `ind.p.3p` and by the plain text definition *Troisième personne du pluriel de l'indicatif présent*. Ideally, the two views of the same fact should be generated from the same data. In other words, the plain text definition should be generated from `ind.p.3p`. Instead, it has been manually typed by a contributor. In this example, the redundant information is consistent. Section 3.2 illustrates situations of inconsistencies.
3. the infrastructure, intended to receive contributions in mass, is in reality restricted to internet users who feel at ease with wikicode editing.

The two first items impact both the quality of Wiktionary itself and the conversion process described in section 3.2. The latter item may lead to an under-participation to the project, and a bias

⁵<http://www.evanjones.ca/software/wikipedia2text.html>

Modes impersonnels				
Mode	Présent		Passé	
Infinitif	affluer	/a.flɥe/	avoir afflué	/a.vwaʁ_a.flɥe/
Gérondif	en affluant	/ɑ̃.n_a.flɥɑ̃/	en ayant afflué	/ɑ̃.n_ɛ.jɑ̃.t_a.flɥe/
Participe	affluant	/a.flɥɑ̃/	afflué	/a.flɥe/

Indicatif				
Présent			Passé composé	
j'afflue	/ʒ_a.flɥ/		j'ai afflué	/ʒ_e a.flɥe/
tu afflues	/ty a.flɥ/		tu as afflué	/ty a.z_a.flɥe/
il/elle/on afflue	/[il/ɛl/ɔ̃] a.flɥ/		il/elle/on a afflué	/[i.l/ɛ.l/ɔ̃.n]_a.a.flɥe/
nous affluons	/nu.z_a.flɥɔ̃/		nous avons afflué	/nu.z_a.vɔ̃.z_a.flɥe/
vous affluez	/vu.z_a.flɥe/		vous avez afflué	/vu.z_a.ve.z_a.flɥe/
ils/elles affluent	/[il/ɛl].z_a.flɥ/		ils/elles ont afflué	/[i/ɛ].z_ɔ̃.t_a.flɥe/
Imparfait			Plus-que-parfait	
j'affluais	/ʒ_a.flɥɛ/		j'avais afflué	/ʒ_a.vɛ.z_a.flɥe/
tu affluais	/ty a.flɥɛ/		tu avais afflué	/ty a.vɛ.z_a.flɥe/
il/elle/on affluait	/[il/ɛl/ɔ̃] a.flɥɛ/		il/elle/on avait afflué	/[i.l/ɛ.l]
nous affluions	/nu.z_a.flɥjɔ̃/			/ɔ̃.n]_a.vɛ.t_a.flɥe/
vous affluiez	/vu.z_a.flɥje/		nous avions afflué	/nu.z_a.vjɔ̃.z_a.flɥe/
ils/elles affluaient	/[il/ɛl].z_a.flɥɛ/		vous aviez afflué	/vu.z_a.vje.z_a.flɥe/
			ils/elles avaient afflué	/[i/ɛ].l.z_a.vɛ.t_a.flɥe/

Fig. 3: Excerpt of the inflectional paradigm of the verb *affluer* in Wiktionnaire

regarding what kind of internet users contribute to Wiktionary. A good initiative, first appeared as an optional *gadget* (in Wiktionary's jargon), is the input field designed to add translations: once a contributor has typed a translation, the graphical interface carries out the corresponding edition of the wikicode. Thus, users unable to edit the wikicode can contribute, and the interface generates an error-free encoding.

The wikicode is volatile over time and is unstable from a language edition to the other. Thus, a parser written for a given edition has to be maintained and cannot be used without adaptation to parse another language edition. A direct consequence is that no fully-automatic update of GLAWI is desirable: potential changes in the wikicode have to be monitored to adapt a given parser to every release of a new dump.

“Experts and Crowds” rather than “Experts vs. Crowds”. Like Wikipedia, Wiktionary is a wiki that any internet user willing to contribute can edit, whatever their skills, with immediate effect. Zesch and Gurevych (2010) assessed Wiktionary's usefulness for semantic relatedness computation. Thus, they illustrated the potential of Wiktionary as a resource for NLP, not its primary quality as a dictionary. Kosem et al. (2013) rely on crowdsourcing in a controlled way to per-

form specific tasks: identifying false collocations and incorrect examples among automatically selected ones. The case of Wiktionary is different: the resource is entirely crowdsourced, with no strong editorial constraint. The legitimacy of the so-called “*wisdom of crowds*” in a lexicographical perspective is discussed by Penta (2011) and Sajous et al. (2014). Regarding Wiktionnaire, it is worth noting that a binary opposition between experts and crowds is not accurate because it has been primarily bootstrapped by automatic imports from editions of two dictionaries fallen into the public domain. Table 1 shows that more than 16% of the entries corresponding to lemmas originate from the 8th edition (1932-1935) of the *Dictionnaire de l’Académie française* (DAF8) or from the 2nd edition (1872-1877) of the *Littré*. The table also reports the number of articles that refer to another resource (only resources with more than 100 references are listed).⁶ These resources include public-domain editions of digitized dictionaries (DAF8, Littré, Bescherelle, Rivarol), Latin (Gaffiot) or Provençal (Mistral) dictionaries, institutional normative websites such as FranceTerme (France) and GDT (Quebec) and specialized websites (Meyer, an online dictionary of animal sciences).

# Imports	# Articles	Percentage
0	242499	83.42%
1	48162	16.57%
2	46	0.02%

Import sources	# Articles	Percentage
DAF8	27945	57.91%
Littré	20278	42.02%
Larousse XIXe	24	0.05%

# References	# Articles	Percentage
0	260362	89.56%
1	27818	9.57%
2	2268	0.78%
3	208	0.07%
4	32	0.01%

Reference sources	# Articles	Percentage
Littré	6497	19.56%
DAF8	6311	19.00%
TLFi	6256	18.84%
Rivarol	4358	13.12%
Meyer	3523	10.61%
FranceTerme	2922	8.80%
Mistral	650	1.96%
ODS5	394	1.19%
GDT	200	0.60%
DAF9	195	0.59%
Bescherelle	116	0.35%
Gaffiot	105	0.32%
Reverso	100	0.30%

Table 1: Imports and references in Wiktionnaire’s articles (lemmas)

⁶A reference means that a contributor manually indicated that she/he consulted a given resource when editing an article.

3. GLAWI

3.1 Resource description

GLAWI is a machine-readable dictionary resulting from the conversion of the Wiktionnaire into an XML-structured format. The resource, released under a free license (CC By-SA),⁷ contains 1,341,410 articles, one for each page of Wiktionnaire. GLAWI's general structure is similar to that to Wiktionnaire's one as exemplified by the article of *mousse* given in Figure 4.

```
<article>
  <title>mousse</title>
  <pageId>7930</pageId>
  <meta>
    <category>Lexique en français de la navigation</category>
    <category>Noms multigenres en français</category>
    <reference>TLFi</reference>
  </meta>
  <text>
    <pronunciations>
      <pron region="France">mus</pron>
    </pronunciations>
    <pos type="nom" lemma="1" locution="0" homoNb="1" gender="f" number="s">
      <paradigm>
        <wiki>{{fr-rég|mus}}</wiki>
        <inflection form="mousse" gracePOS="Ncfs" pron="mus"/>
        <inflection form="mousses" gracePOS="Ncfp" pron="mus"/>
      </paradigm>
      ...
    </pos>
    <pos type="nom" lemma="1" locution="0" homoNb="2" gender="m" number="s">
      ...
    </pos>
    <pos type="nom" lemma="1" locution="0" homoNb="3" gender="m" number="s">
      ...
    </pos>
    <pos type="adjectif" lemma="1" locution="0" gender="epicene" number="s">
      ...
    </pos>
    <pos type="verbe" lemma="0" locution="0">
      <inflectionInfos>
        <inflected gracePOS="Vmip1s-" lemma="mousser" pron="mus"/>
        <inflected gracePOS="Vmip3s-" lemma="mousser" pron="mus"/>
        <inflected gracePOS="Vmisp1s-" lemma="mousser" pron="mus"/>
        <inflected gracePOS="Vmisp3s-" lemma="mousser" pron="mus"/>
        <inflected gracePOS="Vmmp2s-" lemma="mousser" pron="mus"/>
      </inflectionInfos>
    </pos>
  </text>
</article>
```

Fig. 4: General structure of an article in GLAWI: *mousse* entries

⁷GLAWI is available at <http://redac.univ-tlse2.fr/lexicons/glawi.html>

The meta section. The **meta** markup is used to indicate that an article has been imported from, or refers to another dictionary (cf. section 2): the article *nénuphar* (Figure 5) has been primarily imported from DAF8, while the article *mousse* (Figure 4) refers to the TLFi.

```
<article>
  <title>nénuphar</title>
  <pageId>168915</pageId>
  <meta>
    <category>Plantes en français</category>
    <category>Fleurs en français</category>
    <import>DAF8</import>
    <spellingVariation norm="nénufar"/>
  </meta>
  ...
</article>
```

Fig. 5: GLAWI’s metadata for article *nénuphar*

This same section is also used to indicate that an article corresponds to a spelling variant such as *nénuphar*, an alternative form of *nénufar*. Just as in Wikipedia, categories are assigned to pages in Wiktionary. GLAWI’s **meta** section indicates the categories an article belongs to (if any): for example, *mousse* belongs to nautical slang and is a multigender noun ; *nénuphar* belongs to the *Flowers* and *Plants* categories.

POS sections. Articles may contain several POS sections marked by **pos** tags that include grammatical features such as gender, number, valency, homograph number (when relevant) and specify whether a form is multiword or not. An attribute also indicates the lemma of the inflected forms. For example, in Figure 4, the *verb* **pos**-section specifies that *mousse* corresponds to five inflected forms of the verb *mousser* and gives their morphosyntactic descriptions in GRACE format (Rajman et al., 1997).

POS sections also include translations, lexical semantic (synonyms, antonyms, hypernyms, etc.) and morphological (derivative, compound, etc.) relations. An example of such subsections is given in Figure 6 for the feminine noun *mousse* ‘foam’, ‘moss’.

Definitions. Word senses, marked by **definition** tags, are listed in the POS sections of lemmas. A definition contains a gloss and possibly one or more usage examples. Definitions may include labels that give attitudinal, diatopic, diachronic, diafrequential information or indicate that the word belongs to a specialized language. The example in Figure 7 indicates that *mousse*, when used to refer to a beer, is a familiar metonym.

```

<pos type="nom" lemma="1" locution="0" homoNb="1" gender="f" number="s">
...
<subsection type="translations">
  <trans lang="de">Moos</trans>
  <trans lang="en">foam</trans>
  <trans lang="en">moss</trans>
  <trans lang="es">espuma</trans>
</subsection>
<subsection type="lexSemRel">
  <item type="synonym">écume</item>
  <item type="synonym">bière</item>
</subsection>
<subsection type="morphoRel">
  <item type="derivative">moussant</item>
  <item type="derivative">mousser</item>
  <item type="derivative">mousseux</item>
</subsection>
...
</pos>

```

Fig. 6: GLAWI's lexical relations: translations, lexical semantic and morphological relations

```

<pos type="nom" lemma="1" locution="0" homoNb="1" gender="f" number="s">
...
<definitions>
  <definition>
    <gloss>
      <labels>
        <label type="sem" value="métonymie"/>
        <label type="attitudinal" value="familier"/>
      </labels>
      <wiki>{{méton|fr}} {{familier|fr}} [[bière|Bière]].</wiki>
      <txt>Bière.</txt>
      <xml><innerLink ref="bière">Bière</innerLink>.</xml>
      <conll>1 Bière  bière  NC    nc    n=s      0    root</conll>
    </gloss>
    <example>
      <wiki>''Une bonne '''mousse''' sans faux-col est un oxymore.'''</wiki>
      <xml><i>Une bonne <b>mousse</b> sans faux-col est un oxymore.</i></xml>
      <txt>Une bonne mousse sans faux-col est un oxymore.</txt>
      <conll> 1 Une      une      DET  DET  g=f|n=s      3    det
              2 bonne   bon      ADJ  adj  g=f|n=s      3    mod
              3 mousse  mousse  NC   nc   n=s          6    suj
              4 sans    sans    P    P    -            3    dep
              5 faux-col _      NC   -    -            4    prep
              6 est     être    V    v    n=s|p=3|t=pst 0    root
              7 un      un      DET  DET  g=m|n=s      8    det
              8 oxymore _      NC   -    -            6    ats</conll>
    </example>
  </definition>
  ...
</definitions>
...
</pos>

```

Fig. 7: A given sense of *mousse* (feminine noun, homograph #1) as a metonym for *bière* ‘bier’

This figure also shows that every textual part (gloss, example) is available in four different versions:

1. the original wikicode;
2. an XML formatted version where markups encode wiki typesetting (boldface, italic, etc.), dates, foreign words, mathematical/chemical formulae and external/inner links;
3. a raw text version;
4. a CoNLL (Nivre et al., 2007) output of the Talismane syntactic parser.

The XML version of the textual parts could be used to generate other customized versions of the definitions or the etymology sections. The relevance of some elements is actually task-dependent: markups can be used for example to remove non-textual content (formulae) or unwanted words (foreign words). Links can be used by a weighting scheme in information retrieval (Cutler et al., 1997) or to build hyperlink graphs for semantic similarity computation (Weale et al., 2009). The original format is intended for developers that need specific extractions or conversions. Parsed definitions can have various use. Hathout et al. (2014a) for example, leveraged them to acquire morphological relations.

Phonemic transcriptions. 94% of GLAWI's entries contain one or several phonemic transcriptions, potentially including diatopic variations. A given transcription may occur at the article level, and therefore correspond to all the forms described in the article. Transcriptions may also appear in POS sections, especially when homographs have different pronunciations. Figure 8 shows two **pos**-sections of two homographs of *plus*, both adverbs (other POSs omitted). The first one, used in affirmative clauses, is a superlative or a comparative pronounced /ply/ or /plys/. The second homograph, used in negative clauses, is pronounced /ply/.

```
<text>
...
<pos type="adverbe" lemma="1" locution="0" homoNb="1" gracePOS="Rgp">
  <pronunciations>
    <pron>ply</pron>
    <pron>plys</pron>
  </pronunciations>
...
</pos>
<pos type="adverbe" lemma="1" locution="0" homoNb="2" gracePOS="Rgp">
  <pronunciations>
    <pron>ply</pron>
  </pronunciations>
...
</pos>
...
</text>
```

Fig. 8: Phonemic transcriptions of *plus*

In Figure 9, the transcriptions for *moins*, given at the entry level, indicate that for all parts of speech, *moins* is pronounced /mwẽ/ both in “standard” French (Paris) and /mwẽs/ in Southern France (Marseille, Haut Languedoc).

```
<text>
...
<pronunciations>
  <pron region="Marseille, Haut Languedoc">mwẽs</pron>
  <pron region="Paris, France">mwẽ</pron>
</pronunciations>
...
</text>
```

Fig. 9: Phonemic transcriptions of *moins*

3.2 Conversion process: the boundary between standardizing and correcting

As aforementioned, a significant contribution of GLAWI is the standardization of Wiktionnaire’s microstructure⁸ where a given type of information may appear under different forms (predefined templates, aliases, hardcoded text typed by contributors, etc.), and where the same piece of information appearing at different places may lead to inconsistencies. We present two representative examples of consistency checks and standardizing which illustrate the boundary between standardizing and correcting.

Linguistic labels. Contributors can use predefined templates to attach linguistic labels to given definitions. Unlike the English Wiktionary where only two templates (context and label), apparently interchangeable, are used to introduce all the linguistic labels (e.g. {{label|dated}}, {{label|transitive}}, {{label|oenology}}), Wiktionnaire has no generic prefix for these labels: {{désuet}}, {{transitif}} and {{oenologie}}. Detecting linguistic labels in definitions is an important step:

1. to remove them from definitions in order to obtain “clean” text ;
2. to encode the labels into formal markups to ease look-ups (e.g. to target a given label).

Processing the large number of labels used in Wiktionnaire is made even more difficult by their numerous aliases. The diachronic label {{vieilli}} ‘old’, for instance, also occurs under the forms {{vieux}} and {{vx}}. The domain label {{oenologie}} has three other aliases {{œnologie}} (ligature), {{œnol}} and {{œnol}} (abbreviations). A contributor may also ignore these templates and type the domain name between brackets (*oenologie*) directly in the

⁸Complementary details on the extraction process required to convert Wiktionnaire’s loosely wiki-encoded data into a structured format can be found in (Navarro et al., 2009; Sajous et al., 2013b; Hathout et al., 2014a).

definition. We inventoried more than 6,000 different labels and aliases used in definitions to normalize the different ways the same information is encoded. As there is no reason to expect that linguistic labels are used in a more relevant (or, at least, coherent) way in Wiktionnaire than in experts-written dictionaries (Baider et al., 2011), we made no attempt to normalize them further. However, we grouped the linguistic labels into categories (*diatopic*, *diachronic*, *attitudinal*, etc.) that are not encoded in Wiktionnaire. A help page⁹ enumerates most of the labels and classifies them into (questionable) categories: *anglicisme*, *germanisme* and *hispanisme* for example, fall into the *registres d’emploi* ‘usage registers’ category, just as *désuet* ‘obsolete’, *rare* ‘rare’ or *enfantin* ‘childish’ do. The label *euphémisme* (euphemism) appears under the category *relations entre les sens* ‘relations between senses’ whereas *dérision* ‘derision’, *mélioratif* ‘meliorative’ and *péjoratif* ‘pejorative’ belong to *registres d’emploi*. This latter category contains the label *informel* ‘informal’ while *soutenu* ‘formal’ belongs to *registres de langue* ‘level of language’. We did not use these categories and decided to manually build coarse-grained ones to which each label can be assigned. Except for the aforementioned normalization of aliases, we did not modify label values and maintained label pairs that look interchangeable. For example, if the difference between *archaïque* ‘archaic’ and *vieilli* ‘old’ is clear, *vieilli* and *désuet* are not clearly distinguished:

- *désuet* = “pour indiquer que le mot vedette n’est plus employé par la langue moderne” ‘to indicate that a headword is not used any longer in modern language’
- *vieilli* = “pour indiquer que le mot vedette est vieilli” ‘to indicate that a headword is dated’

Similarly, guidance could be expected to differentiate *littéraire* from *soutenu*, but *littéraire* has no definition and the use of *soutenu* is recommended when the headword belongs to the language level... *soutenu*.

Inflectional paradigms. We have described Wiktionnaire’s macrostructure in section 2 and shown the multiple links between the paradigm of a lemma and the corresponding inflected forms. The four inflected forms of the adjective *affluent* (Fig. 1a) are generated by the wiki template `{{fr-accord-cons|a.fly.â|t}}` (Fig. 1b). Parsing the article dedicated to the form *affluente* (Fig. 1c) confirms that it is the feminine singular form of the adjective *affluent*. However, scattered information is not always redundant: for instance, the gender of the noun *arrivages* ‘arrivals’ is missing in the corresponding page;¹⁰ but the definition indicates that this entry is the plural of *arrivage* ‘arrival’. The masculine gender of *arrivage* being mentioned in its page, we can infer that *arrivages* is masculine too. Unfortunately, contradictory information occurs as well. For example, in the page *clavardeuses*¹¹ (chatters, feminine plural noun in French from Quebec), the gender of the entry is specified as *masculine* whereas the definition states “*Féminin pluriel de clavardeur*”. In such cases, information is left as is and an “*inconsistent*” attribute is added to the GLAWI’s entry (only 65 entries are concerned).

⁹http://fr.wiktionary.org/wiki/Wiktionnaire:Liste_de_tous_les_modèles/Précisions_de_sens

¹⁰<http://fr.wiktionary.org/w/index.php?title=arrivages&oldid=19099721>

¹¹<http://fr.wiktionary.org/w/index.php?title=clavardeuses&oldid=19129490>

All the inflectional information is propagated in this way and if some features are still missing, we lookup in Lefff (Sagot et al., 2006) and Morphalou (Romary et al., 2004) to fill some of the lacks. We used these lexicons to complete GLAWI by adding:

- 366 missing lemmas of inflected forms having full morphosyntactic description in Wiktionnaire;
- 17,446 incomplete morphosyntactic description of inflected forms whose lemma is known;
- 444 genders of nouns or adjectives.

After this last completion, 1.4% of the inflected adjectival forms and 3.7% of the inflected nominal forms still have a missing number or gender (when considering monolexical forms only).

Verb paradigms may be problematic as well: missing inflected forms may be lacking or denote verb defectiveness. Several forms for a given inflection may originate from a superabundant verb, or results from inconsistencies. For example, the conjugation page of *payer*¹² ‘to pay’ gives the two paradigms of this verb. An apparently similar case could explain the two forms *contredisez* and *contredites* of the second person plural of the verb *contredire* ‘to contradict’, imperative mood. The former is the correct form, found in the corresponding page. The latter, given in the conjugation table¹³, is erroneous. Another example is given by the two forms *végèterai/végerai* of the verb *végéter* ‘to vegetate’, first person singular of future indicative, which are neither erroneous nor superabundant. The former is the modern spelling while the latter corresponds to the spelling in use before the 1976 orthographic reform. This latter case is easy to deal with as a specific template identifies the *é/è* alternations due to this reform. In such case, the detected phenomenon is reported into GLAWI by a specific markup. When there is no element to decide whether forms are legitimate or erroneous, we include them all, leaving the opportunity to the users exploiting GLAWI to perform subsequent processing. Handling such cases can also constitute a possible improvement for future versions of GLAWI.

3.3 Next steps

From GLAWI back to Wiktionnaire? GLAWI’s existence is only possible thanks to the contributions of the wiktionarians. Reciprocally, the efforts we made in the standardization and consistency checking process could benefit Wiktionnaire, even if the collaboration between academics and wiktionarians may not be self-evident. Wikis are sometimes presented as knowledge democracy. Hanks (2012) presents Wiktionary as an “*anarcho-syndicalist approach to lexicography*”; Meyer and Gurevych (2012) write that Wiktionary is constructed by a large community of ordinary web users and that the community has a lively discussion culture. In reality, the community only has a small number of *active* contributors who perform most of the contributions: only 117 contributors to Wiktionnaire performed at least 5 edits in March 2015; 35 of

¹²http://fr.wiktionary.org/wiki/Annexe:Conjugaison_en_français/payer

¹³http://fr.wiktionary.org/w/index.php?title=Annexe:Conjugaison_en_français/contredire&oldid=8789428

them performed at least 100 edits.¹⁴ These contributors often have responsibility in the management of the dictionary: each wiki project functions as an ecosystem with its administrators, patrollers, functionaries, clerks, bots, etc. There is no denying that discussions may be lively, but they essentially take place among the small world of active contributors. The observation of Wiktionnaire's discussion pages shows that hours of voluntary work make the contributors quite reluctant to be "dispossessed" from the fruits of their labor. In this context, a newcomer, whether or not a language professional, has to become part of the community before getting credit and fruitfully proposing changes. Anyway, we will not seek to impose standardization or corrections. We take Wiktionary as it is: Wiktionnaire would certainly have attracted fewer contributors if it was more constrained. GLAWI is at the wiktionarians' disposal, who can use it to reinject information in Wiktionnaire if the community judge it relevant.

Forward synchronization. We previously mentioned Wiktionary's potential for constant update. We also highlighted that its volatile format makes regular fully-automatic conversions impossible. In order to reflect Wiktionnaire's up-to-dateness, new versions of GLAWI will be released in the future. GLAWI update frequency will however not follow the periodicity of XML dumps releases: manual checks have to be performed to ensure that a given parser is still compliant with a new dump. If not, maintenance is required to adapt to format changes.

Other languages. Similarly, due to the format heterogeneity between all language editions, adapting a parser designed for a given language to another one may require heavy changes. Hence, the benefits that can be expected from such work have to be balanced with the size of the targeted language edition and its estimated quality/density. Regarding the size, the number of articles per edition ranges from 45 to more than 4 million¹⁵ and is not necessarily correlated with the number of native speakers: for instance, the second most represented language in Wiktionary is Malagasy while (Mandarin) Chinese ranks sixth.

4. From GLAWI to on demand tailored lexicons

GLAWI has been used to create a number of customized lexicons dedicated to specific uses including NLP, linguistic description and psycholinguistics. The main one is GLÀFF, a large inflectional and phonological lexicon of French. We also derived from GLAWI a morphological derivational resource and a list of humans names.

GLÀFF, a large inflectional and phonological lexicon of French. Collecting the inflectional and phonological information described in GLAWI is quite easy. We just need to traverse the XML file and fill them into the lexicon slots. Since GLAWI provides morphosyntactic tags, we do not even have to parse the inflected words definitions nor the inflectional paradigms of

¹⁴<http://stats.wikimedia.org/wiktionary/EN/TablesRecentTrends.htm>

¹⁵The number of articles per language edition is given at: https://meta.wikimedia.org/wiki/Wiktionary#List_of_Wiktionaries

the lemmas. Similarly, GLAWI makes the phonological information available in API with the syllables boundaries. No further processing is needed to fill in the phonological fields in the lexicon.

The extracted lexicon called GLÀFF includes more than 1.4 million entries, each one containing a wordform, a tag in GRACE format, a lemma and, when present in Wiktionnaire, phonemic transcriptions (cf. Fig. 10). Entries also contain word frequencies computed over different corpora.

affluent Afpms affluent a.fly.ɑ̃ 12 0.41 15 0.51 175 0.79 183 0.83 576 0.45 696 0.55
affluente Afpfs affluent a.fly.ɑ̃t 0 0 0 0 2 0.00 183 0.83 9 0.00 696 0.55
affluents Afpmp affluent a.fly.ɑ̃ 2 0.06 15 0.51 5 0.02 183 0.83 89 0.07 696 0.55
affluentes Afpfp affluent a.fly.ɑ̃t 1 0.03 15 0.51 1 0.00 183 0.83 22 0.01 696 0.55
affluent Ncms affluent a.fly.ɑ̃ 22 0.76 38 1.31 232 1.05 444 2.02 1234 0.98 3655 2.91
affluents Ncmp affluent a.fly.ɑ̃ 16 0.55 38 1.31 212 0.96 444 2.02 2421 1.93 3655 2.91
affluent Vmip3p- affluer a.fly 9 0.31 187 6.48 369 1.67 1207 5.49 500 0.39 1929 1.53
affluent Vmisp3p- affluer a.fly 9 0.31 187 6.48 369 1.67 1207 5.49 500 0.39 1929 1.53

Fig. 10: Extract of GLÀFF



GLÀFF is by far larger than any other inflectional and/or phonological lexicon of French we know of. Sajous et al. (2013a), Hathout et al. (2014b) and Sajous et al. (2014) compare GLÀFF with four of them¹⁶ and show that it contains 3 to 4 times more lemmas and 3 to 9 times more inflected forms. This size is an important asset when the lexicon is used for research in derivational or inflectional morphology. It is also an advantage for the development of NLP tools such as morphosyntactic taggers and parsers. The comparison also reveals that GLÀFF has a better coverage of the vocabulary of corpora of various types and that it includes many usual words such as: *attractivité* ‘attractivity’, *diabolisation* ‘demonetization’, *homophobie* ‘homophobia’ or *hébergeur* ‘host’, etc. missing from the other lexicons. In addition, GLÀFF’s phonemic transcriptions are highly consistent with those of BDLex and Lexique.

Another interesting feature of GLÀFF is its online browsing interface, called GLÀFFOLI.¹⁷ This interface, illustrated in Figure 11, enables any user to build a multicriteria query. Request fields may include wordform, lemma, part of speech and/or pronunciation. When the user chooses to display corpora frequencies, the wordforms attested in FrWaC are linked to the NoSkecthEngine concordancer (Rychlý, 2007).

PsychoGLÀFF. GLÀFF has in turn been used to create an even more specific lexicon designed to meet the psycholinguistic needs. Calderone et al. (2014) present PsychoGLÀFF, a version of GLÀFF especially dedicated to the creation and calibration of experimental material that provides a range of additional features of the phonological and written forms such as frequency, lexical neighborhoods, syllabic complexity and phonotactic likelihood.

¹⁶The aforementioned morphological lexicons Lefff and Morphalou ; Lexique (New, 2006), a free lexicon popular in psycholinguistics, which contains phonemic transcriptions but has a restricted coverage ; BDLex (Pérennou and de Calmès, 1987) a non-free lexicon with both an exploitable coverage and phonemic transcriptions.

¹⁷<http://redac.univ-tlse2.fr/glaffoli/>


[\[Search \]](#) [\[Preferences \]](#) [\[About/Help \]](#)


Form IPA

Lemma

POS SAMPA

☐ attested forms only

Reset Search

Showing results 1-100 of 235

[\[Export as CSV \]](#)

Form	POS	Lemma	IPA	SAMPA	Frantext 20 ^e		LM10		FrWaC	
					Form ↓ ↑	Lemma ↓ ↑	Form ↓ ↑	Lemma ↓ ↑	Form ↓ ↑	Lemma ↓ ↑
achilletalonesques	Afpfp	achilletalonesque	a.ʃil.ta.lɔ̃.nɛsk	a.sil.ta.lɔ̃.nɛsk	0 0	0 0	0 0	0 0	0 0	0 0
capillaires	Afpfp	capillaire	ka.pi.lɛʁ	ka.pi.lɛʁ	12 0.415	20 0.693	87 0.395	144 0.655	1123 0.895	3019 2.407
capillotractées	Afpfp	capillotracté	ka.pi.lɔ̃.tʁak.te	ka.pi.lɔ̃.tʁak.te	0 0	0 0	0 0	0 0	2 0.001	11 0.008
bacillaire	Afpfs	bacillaire	ba.si.lɛʁ	ba.si.lɛʁ	1 0.034	2 0.069	2 0.009	2 0.009	36 0.028	44 0.035
ancillaire	Afpfs	ancillaire	ɑ̃.si.lɛʁ	A~.si.lɛʁ	10 0.346	25 0.866	10 0.045	25 0.113	66 0.052	128 0.102

Fig. 11: GLÀFFOLI, the GLÀFF OnLine Interface

Extracting derivational relations from GLAWI. GLAWI actually provides information on all aspects of morphology including derivational morphology. Hathout et al. (2014a) present several methods to acquire derivational relations and morpho-semantic knowledge. The first is simply to extract the derivational relations listed in GLAWI's morphoRel tags. A second, and more sophisticated method, acquires the relations from the morphological definitions, that is, definitions where the *definiens* contains a word from the morphological family of the *definiendum*. These relations were then further filtered out so that only the ones that can form analogies with the relations listed in morphoRel tags were kept. Over all, the derivational resource that resulted from this acquisition contains more than 170,000 relations and is the largest one available for French at the moment.

Human names extraction. Flaux et al. (2014) study the human names that denote a creative activity, such as *symphoniste* (symphonist), *sculpteur* (sculptor) or *romancier* (novelist). Such names have been collected into the NHUMA database¹⁸ from different sources such as a language dictionary (TLFi), a dictionary of synonyms (DicoSyn) and WaliM (Namer, 2003), a tool for harvesting the web. After these resources have been exploited, a simple lookup in GLAWI's glosses, based on lexical cues only, enabled a 15% increase of the database.

Other possibilities. Filtering GLAWI's linguistic labels or other markups instantly permits on demand tailoring of lexicons such as loanwords used in French, masculine/feminine noun equivalents, dated words, domain-specific sublexicons, etc. Regarding lexicography, an immediate application could be the use of GLAWI for neology monitoring. Automatic detection of neologisms in corpora produces a lot of noise. GLAWI can be used to detect true positives among the candidates. When a form extracted from a corpus is absent from the reference lexicon, its occurrence in GLAWI is a serious hint of actual neology.

¹⁸<http://nomsdhumains.weebly.com>

5. Conclusion and perspectives

This paper introduces GLAWI, an XML-encoded machine-readable dictionary automatically extracted from Wiktionnaire. Therefore, GLAWI inherits most of Wiktionnaire's strong points, including the exceptional number of its headwords and an original macrostructure. This has been assessed through detailed comparisons with well-known inflectional and phonological lexicons.

Wiktionnaire's editorial success is linked to its use of MediaWiki which imposes no constraint on how information is represented. The flip side is the great heterogeneity of its microstructure which makes it difficult to use in NLP and prevents the selection of articles with targeted queries such as "I am looking for particle nouns ending in *-on*" like *neutron*, *gluon* or *boson*. GLAWI specifically addresses these needs: the XML markups encode the microstructure explicitly; it standardizes the Wiktionnaire's content and enhances its coherence, standardization being clearly a prerequisite to any automated exploitation.

GLAWI is also an answer to other needs, like the creation of specific lexical resources. Indeed, it is likely that the development of the mobile web is changing the way users access MRDs. Complex interfaces like the one of the *Trésor de la Langue Française informatisé* (TLFi), a large French MRD (Dendien, 1994), are losing ground in favor of applications built around specific information subsets such as thesauri, quotation, slang, rhyming, etymological or bilingual dictionaries, but also less traditional derivative works like dictionaries of Latin loanwords, morphological dictionaries or dictionaries of epicene nouns. However, the need to access dictionaries through targeted queries remains, particularly for skilled users (Lew, 2013) and for language specialists, especially linguists and lexicographers. To this end, we plan to design a user-friendly interface for GLAWI, similar to GLÀFFOLI (see Figure 11).

Another remarkable feature GLAWI inherits from Wiktionnaire is its free license which makes it a resource adapted to current research practice in NLP. NLP is indeed becoming a discipline where experimentation occupies an increasingly important place and where experiment replication is becoming common. One consequence of this development is the requirement to use freely available resources and data sets. GLAWI fulfills this condition but similar resources for French are in short supply as traditionally, researchers and labs greatly restrict the access to the data they produce. Notable exceptions are Lefff, an inflectional lexicon used by several taggers, Lexique, until recently the only free resource including phonemic transcriptions and Flexique (Bonami et al., 2014), produced by semi-automatically filling the paradigms of Lexique's entries. Notice however that there is no satisfactory resource providing definitions. TLFi is not available for download and, according to Eckard et al. (2012), WOLF (Sagot and Fišer, 2008), a free French WordNet built automatically by aggregating and translating other resources, is sparse and not completely translated. The lack of free satisfactory lexical resources does not only impact research. It is also an impediment to the development of language processing applications. The long-term survival of dictionaries is questioned by Rundell (2012), who envisages that their heterogeneous functions might be better performed by separate specialized tools. If this happens, such tools, while contributing to the disappearance of dictionaries in their current forms, will still necessitate lexical knowledge embedded in electronic dictionaries. GLAWI could meet such needs.

Acknowledgments

The authors would like to thank the anonymous reviewers for their insightful comments.

Syntactic parsing has been performed using the OSIRIM platform that is administered by IRIT and supported by CNRS, the Region Midi-Pyrénées, the French Government and ERDF.

Bibliography

- Baider et al., 2011. Baider, F., Lamprou, E., and Monville-Burston, M. (2011). *La marque en lexicographie: états présents, voies d'avenir*. La lexicothèque. Lambert-Lucas.
- Bonami et al., 2014. Bonami, O., Caron, G., and Plancq, C. (2014). Construction d'un lexique flexionnel phonétisé libre du français. In *Actes du 4^e Congrès Mondial de Linguistique Française (CMLF 2014)*, pages 2583–2596, Berlin.
- Calderone et al., 2014. Calderone, B., Hathout, N., and Sajous, F. (2014). From GLÀFF to PsychoGLÀFF: a large psycholinguistics-oriented French lexical resource. In *Proceedings of the 16th EURALEX International Congress*, pages 431–446, Bolzano.
- Calzolari, 1988. Calzolari, N. (1988). The dictionary and the thesaurus can be combined. In Evens, M., editor, *Relational Models of the Lexicon*, pages 75–96. Cambridge University Press.
- Chodorow et al., 1985. Chodorow, M. S., Byrd, R. J., and Heidorn, G. E. (1985). Extracting semantic hierarchies from a large on-line dictionary. In *Proceedings of the 23rd Annual Meeting on Association for Computational Linguistics, ACL'85*, pages 299–304, Chicago.
- Cutler et al., 1997. Cutler, M., Shih, Y., and Meng, W. (1997). Using the Structure of HTML Documents to Improve Retrieval. In *Proceedings of the USENIX Symposium on Internet Technologies and Systems*, pages 241–252, Monterey.
- Dendien, 1994. Dendien, J. (1994). Le projet d'informatisation du TLF. In Éveline Martin, editor, *Les textes et l'informatique*, chapter 3, pages 31–63. Didier Érudition, Paris, France.
- Eckard et al., 2012. Eckard, E., Barque, L., Nasr, A., and Sagot, B. (2012). Dictionary-Ontology Cross-Enrichment. Using TLFi and WOLF to enrich one another. In *COLING Workshop on Cognitive Aspects of the Lexicon*, pages 81–93, Mumbai.
- Flaux et al., 2014. Flaux, N., Lagae, V., and Stosic, D. (2014). Romancier, symphoniste, sculpteur : les noms d'humains créateurs d'objets idéaux. In *Actes du 4^{eme} Congrès Mondial de Linguistique Française (CMLF 2014)*, pages 3075–3089, Berlin.
- Hanks, 2012. Hanks, P. (2012). Corpus evidence and electronic lexicography. In Granger, S. and Paquot, M., editors, *Electronic Lexicography*, chapter 4, pages 57–82. Oxford University Press, Oxford.
- Hathout, 2011. Hathout, N. (2011). Morphonette: a paradigm-based morphological network. *Lingue e linguaggio*, 2011(2):243–262.
- Hathout and Namer, 2014. Hathout, N. and Namer, F. (2014). Démonette, a French derivational morpho-semantic network. *Linguistic Issues in Language Technology*, 11(5):125–168.
- Hathout et al., 2014a. Hathout, N., Sajous, F., and Calderone, B. (2014a). Acquisition and enrichment of morphological and morphosemantic knowledge from the French Wiktionary. In *Proceedings of the COLING Workshop on Lexical and Grammatical Resources for Language Processing*, pages 65–74, Dublin.
- Hathout et al., 2014b. Hathout, N., Sajous, F., and Calderone, B. (2014b). GLÀFF, a Large Versatile French Lexicon. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC'14)*, pages 1007–1012, Reykjavik.
- Kosem et al., 2013. Kosem, I., Gantar, P., and Krek, S. (2013). Automation of lexicographic work: An opportunity for both lexicographers and crowd-sourcing. In *Proceedings of eLex 2013*, pages 32–48, Tallinn.

- Lew, 2013. Lew, R. (2013). Online dictionary skills. In *Proceedings of eLex 2013*, pages 16–31, Tallinn.
- Markowitz et al., 1986. Markowitz, J., Ahlswede, T., and Evens, M. (1986). Semantically significant patterns in dictionary definitions. In *Proceedings of the 24th Annual Meeting on Association for Computational Linguistics*, pages 112–119, New York.
- Meyer, 2013. Meyer, C. M. (2013). *Wiktionary: The Metalexigraphic and the Natural Language Processing Perspective*. PhD thesis, Technische Universität Darmstadt.
- Meyer and Gurevych, 2012. Meyer, C. M. and Gurevych, I. (2012). Wiktionary: A new rival for expert-built lexicons? Exploring the possibilities of collaborative lexicography. In Granger, S. and Paquot, M., editors, *Electronic Lexicography*, chapter 13, pages 259–291. Oxford University Press, Oxford.
- Namer, 2003. Namer, F. (2003). WaliM : valider les unités morphologiquement complexes par le web. In Fradin, B., Dal, G., Kerleroux, F., Hathout, N., Plénat, M., and Roché, M., editors, *Les unités morphologiques. Actes du 3ème Forum de Morphologie.*, pages 142–150, Lille.
- Navarro et al., 2009. Navarro, E., Sajous, F., Gaume, B., Prévot, L., Hsieh, S., Kuo, I., Magistry, P., and Huang, C.-R. (2009). Wiktionary and NLP: Improving synonymy networks. In *Proceedings of the 2009 ACL-IJCNLP Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources*, pages 19–27, Singapore.
- New, 2006. New, B. (2006). Lexique 3 : Une nouvelle base de données lexicales. In *Verbum ex machina. Actes de la 20e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2006)*, pages 892–900, Louvain-la-Neuve.
- Nivre et al., 2007. Nivre, J., Hall, J., Kübler, S., McDonald, R., Nilsson, J., Riedel, S., and Yuret, D. (2007). The CoNLL 2007 Shared Task on Dependency Parsing. In *Proceedings of the CoNLL 2007 Shared Task on dependency parsing (EMNLP-CoNLL)*, pages 915–932, Prague.
- Penta, 2011. Penta, D. J. (2011). The wiki-fication of the dictionary: defining lexicography in the digital age. In *Proceedings of the MIT7 Conference "unstable platforms: the promise and peril of transition"*, Cambridge.
- Pérennou and de Calmès, 1987. Pérennou, G. and de Calmès, M. (1987). BDLEX lexical data and knowledge base of spoken and written French. In *Proceedings of the European Conference on Speech Technology, ECST 1987*, pages 1393–1396, Edinburgh.
- Rajman et al., 1997. Rajman, M., Lecomte, J., and Paroubek, P. (1997). Format de description lexicale pour le français. Partie 2 : Description morpho-syntaxique. Technical report, EPFL & INaLF. GRACE GTR-3-2.1.
- Romary et al., 2004. Romary, L., Salmon-Alt, S., and Francopoulo, G. (2004). Standards going concrete : from LMF to Morphalou. In *Proceedings of COLING 2004: Enhancing and using electronic dictionaries*, pages 22–28, Geneva.
- Rundell and Kilgarriff, 2011. Rundell, M. and Kilgarriff, A. (2011). Automating the creation of dictionaries: Where will it all end? In Meunier, F., De Cock, S., Gilquin, G., and Paquot, M., editors, *A Taste for Corpora. In honour of Sylviane Granger*, pages 257–282. John Benjamins.
- Rundell, 2012. Rundell, M. (2012). It works in practice but will it work in theory? The uneasy relationship between lexicography and matters theoretical. In *Proceedings of the 15th EURALEX International Congress*, pages 47–92, Oslo.
- Rychlý, 2007. Rychlý, P. (2007). Manatee/Bonito - A Modular Corpus Manager. In *Proceedings of the 1st Workshop on Recent Advances in Slavonic Natural Language Processing*, pages 65–70, Brno.

- Sagot et al., 2006. Sagot, B., Clément, L., De La Clergerie, E., and Boullier, P. (2006). The Lefff 2 syntactic lexicon for French: architecture, acquisition, use. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, pages 1348–1351, Genoa.
- Sagot and Fišer, 2008. Sagot, B. and Fišer, D. (2008). Building a free French wordnet from multilingual resources. In *Proceedings of OntoLex 2008*, Marrakech.
- Sajous et al., 2010. Sajous, F., Navarro, E., Gaume, B., Prévot, L., and Chudy, Y. (2010). Semi-automatic Endogenous Enrichment of Collaboratively Constructed Lexical Resources: Piggybacking onto Wiktionary. In Loftsson, H., Rognvaldsson, E., and Helgadóttir, S., editors, *Advances in Natural Language Processing*, volume 6233 of *LNCS*, pages 332–344. Springer Berlin / Heidelberg.
- Sajous et al., 2013a. Sajous, F., Hathout, N., and Calderone, B. (2013a). GLÀFF, un Gros Lexique À tout Faire du Français. In *Actes de la 20e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2013)*, pages 285–298, Les Sables d'Olonne.
- Sajous et al., 2013b. Sajous, F., Navarro, E., Gaume, B., Prévot, L., and Chudy, Y. (2013b). Semi-automatic Enrichment of Crowdsourced Synonymy Networks: the WISIGOTH System Applied to Wiktionary. *Language Resources and Evaluation, special issue on Collaboratively Constructed Language Resources*, pages 1–34.
- Sajous et al., 2014. Sajous, F., Hathout, N., and Calderone, B. (2014). Ne jetons pas le Wiktionnaire avec l'oripeau du web ! Études et réalisations fondées sur le dictionnaire collaboratif. In *Actes du 4e Congrès Mondial de Linguistique Française (CMLF 2014)*, pages 663–680, Berlin.
- Sérasset, 2012. Sérasset, G. (2012). Dbmary: Wiktionary as a LMF based Multilingual RDF network. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, pages 2466–2472, Istanbul.
- Urieli, 2013. Urieli, A. (2013). *Robust French syntax analysis: reconciling statistical methods and linguistic knowledge in the Talismane toolkit*. PhD thesis, Université de Toulouse II-Le Mirail.
- Weale et al., 2009. Weale, T., Brew, C., and Fosler-Lussier, E. (2009). Using the Wiktionary Graph Structure for Synonym Detection. In *Proceedings of the ACL-IJCNLP Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources*, pages 28–31, Singapore.
- Zesch and Gurevych, 2010. Zesch, T. and Gurevych, I. (2010). Wisdom of Crowds versus Wisdom of Linguists - Measuring the Semantic Relatedness of Words. *Journal of Natural Language Engineering.*, 16(01):25–59.